Incorporating Non-Relevance Information in the Estimation of Query Models

Edgar Meij Wouter Weerkamp Jiyin He Maarten de Rijke

ISLA, University of Amsterdam
http://ilps.science.uva.nl/

Abstract: We describe the participation of the University of Amsterdam's ILPS group in the relevance feedback track at TREC 2008. We introduce a new model which incorporates information from relevant and non-relevant documents to improve the estimation of query models. Our main findings are twofold: (i) in terms of statMAP, a larger number of judged non-relevant documents improves retrieval effectiveness and (ii) on the TREC Terabyte topics, we can effectively replace the estimates on the judged non-relevant documents with estimations on the document collection.

1 Introduction

In our participation in the relevance feedback track this year, our goal was to explicitly incorporate non-relevance information in the estimation of query models. Working with the language modeling approach to information retrieval, we base our model of non-relevant information on the Normalized Log Likelihood Ratio.

We discuss related work in Section 2, describe our retrieval approach in Section 3, and detail our model for capturing non-relevance in Section 4. In Section 5 we describe our runs; we then present our results in Section 6 and conclude in Section 7.

2 Background

Our chief aim for participating in this year's TREC Relevance Feedback track is to extend previous approaches, such as the one proposed by Lavrenko and Croft (2001), by explicitly incorporating non-relevance information. Such negative evidence is usually assumed to be implicit, i.e. in the case of estimating a model from some (pseudo-)relevant data, the absence of terms indicates their non-relevance status. This means, in a language modeling setting and for the sets of relevant documents R and non-relevant documents $\bar{R}$, $P(t|\theta_{\bar{R}}) = 1 - P(t|\theta_R)$. The TREC Relevance Feedback track gives us the opportunity to develop and evaluate models which explicitly capture non-relevance information and we participated to answer the following research questions. Can non-relevance information be effectively modeled to improve the estimation of a query model? Given our model, what is the effect of the relative size of the set of non-relevant documents with respect to the relevant documents on retrieval effectiveness? And, finally, we ask the question whether and when explicit non-relevance information helps. In other words, what are the effects when we substitute the estimates on the non-relevant documents with more general estimates, such as from the collection. Some previous work has already experimented with using negative weights for non-relevance information, either in an ad-hoc or more principled fashion, with mixed results (Dunlop, 1997; Ide, 1971; Wang et al., 2008; Wong et al., 2008).

The model we propose leverages the *distance* between each relevant document and the set of non-relevant documents, by penalizing terms that occur frequently in the latter, similar to the intuitions described by Wang et al. (2008). Instead of subtracting probabilities, however, we take a more principled approach based on the Normalized Log Likelihood Ratio (NLLR). Moreover, similar to other pseudo-relevance feedback approaches, such as the one proposed by Lavrenko and Croft (2001), we reward terms that appear frequently in the individual relevant documents. Although the NLLR is not a true distance between distributions (since it does not satisfy the triangle equality), we consider it to be a useful candidate for measuring the (dis)similarity between two probability distributions.

3 Retrieval Framework

We employ a language modeling approach to IR and rank documents by their log-likelihood of being relevant given a query. Without presenting details here we only provide our final formula for ranking documents, and refer the reader to (Balog et al., 2008) for the steps of deriving this equation:

$$\log P(D|Q) \propto \log P(D) + \sum_{t \in Q} P(t|\theta_Q) \cdot \log P(t|\theta_D). \quad (1)$$

Here, both documents and queries are represented as multinomial distributions over terms in the vocabulary, and are referred to as *document model* (θ_D) and *query model* (θ_Q), respectively. The third component of our ranking model is the *document prior* ($P(D)$), which is assumed to be uniform. Note that by using uniform priors, Eq. 1 gives the same ranking as scoring documents by measuring the KL-divergence between the query model θ_Q and each document model θ_D, in which the divergence is negated for ranking purposes (Lafferty and Zhai, 2001).

Unless indicated otherwise, we estimate each document model by:

$$P(t|\theta_D) = (1-\lambda_D)\cdot P(t|D) + \lambda_D \cdot P(t), \tag{2}$$

where λ_D is a parameter by that we use to tune the amount of smoothing. $P(t|D)$ indicates the maximum likelihood estimate (MLE) of term t on a document, i.e., $P(t|D) = n(t,D)/\sum_{t'} n(t',D)$, and $P(t)$ the MLE on the collection C:

$$P(t) = P(t|C) = \frac{\sum_D n(t,D)}{|C|}. \tag{3}$$

As to the query model θ_Q, we adopt the common approach to linearly interpolate the initial query with an expanded part (Balog et al., 2008; Kurland et al., 2005; Rocchio, 1971; Zhai and Lafferty, 2001):

$$P(t|\theta_Q) = \lambda_Q \cdot P(t|\hat{\theta}_Q) + (1-\lambda_Q)\cdot P(t|Q), \tag{4}$$

where $P(t|Q)$ indicates the MLE on the initial query and the parameter λ_Q controls the amount of interpolation. The main goal of our participation is to find ways of improving the query model $\hat{\theta}_Q$ using (non-)relevance information.

4 Modeling Non-Relevance

Kraaij (2004) defines the NLLR measure as being equivalent to determining the negative KL-divergence for document retrieval. It is formulated as:

$$\mathrm{NLLR}(Q|D) \quad = \quad H(\theta_Q,\theta_C) - H(\theta_Q,\theta_D), \tag{5}$$

where $H(\theta,\theta')$ is the cross-entropy between two multinomial language models:

$$\begin{aligned} H(\theta,\theta') &= H(\theta) + \mathrm{KL}(\theta||\theta') \\ &= -\sum_t P(t|\theta)\log P(t|\theta) + \\ &\qquad \sum_t P(t|\theta)\log\frac{P(t|\theta)}{P(t|\theta')} \\ &= -\sum_t P(t|\theta)\log P(t|\theta'). \end{aligned}$$

Eq. 5 can be interpreted as the relationship between two language models θ_Q and θ_D, normalized by a third language model θ_C (these three models are estimated using Eq. 4, Eq. 2, and Eq. 3 respectively). The NLLR is a measure of average surprise; the better a document model 'fits' a query distribution, the higher the score will be; $H(\theta_Q,\theta_D)$ will be smaller than $H(\theta_Q,\theta_C)$ for relevant documents. In other words, the smaller the cross entropy between the query and document model (i.e., when the document language model better fits the observations from the query language model), the higher it will be ranked.

Based on the NLLR measure, we have developed the following model by which we estimate $P(t|\hat{\theta}_Q)$ in Eq. 4. The intuition is to determine for each term, the probability that it was sampled from each relevant document as well as the probability that it was sampled from the set of non-relevant documents:

$$P(t|\hat{\theta}_Q) \quad \propto \quad \sum_{D\in R} P(t|\theta_D)P(\theta_D|\theta_R),$$

We weigh each term by the distance between R and $\bar{R}$ and its importance in the current document by setting:

$$P(\theta_D|\theta_R) = \frac{\mathrm{NLLR}(D|R)}{\sum_{D'}\mathrm{NLLR}(D'|R)}, \tag{6}$$

where

$$\begin{aligned} \mathrm{NLLR}(D|R) &= H(\theta_D,\theta_{\bar{R}}) - H(\theta_R,\theta_D) \qquad (7) \\ &= \sum_t P(t|\theta_D)\log\frac{P(t|\theta_R)}{P(t|\theta_{\bar{R}})} \\ &= \sum_t P(t|\theta_D)\log\frac{(1-\delta_1)P(t|R)+\delta_1 P(t)}{(1-\delta_2)P(t|\bar{R})+\delta_2 P(t)}. \end{aligned}$$

The δ parameters provide us with the means to control the individual influence of each set of relevant and non-relevant documents versus a background model. $P(t|R)$ and $P(t|\bar{R})$ are estimated by considering the MLE on the documents in the respective set, i.e., for the set of relevant documents R:

$$P(t|R) = \frac{\sum_{D\in R} P(t|D)}{|R|}.$$

5 Runs

We submitted 2 runs, each consisting of 5 separate runs (one for each set of provided relevance judgements). The capital letters in each run indicate the relevance judgements per topic used for that run: (A) no relevance judgements, (B) 3 relevant documents, (C) 3 relevant and 3 non-relevant documents, (D) 10 judged documents (division unknown), (E) large set of judgements (division and number unknown).

We have followed the following intuition for our submissions: given that we have knowledge on which documents are relevant and not relevant to the query, can we use this information to obtain a better estimate of our query model? We hypothesize that our model gains the most when the set of non-relevant documents is large enough to give a proper estimate on non-relevance. We expect the background collection to be a better estimate of non-relevance when the set of

	Set	MAP	P5	P10
	A	0.1364	0.2516	0.2452
met6	B	**0.1732**$^{\triangle}$	0.2645	0.2677
met6	C	0.1568	**0.3484**	**0.3129**
met6	D	0.1584	0.3097	**0.3129**
met6	E	0.1689	0.2645	0.2677
met9	B	0.1769$^{\triangle}$	0.3161	0.3194
met9	C	0.1699$^{\triangle}$	0.3161	0.3032
met9	D	0.1738$^{\triangle}$	**0.4000**$^{\triangle}$	**0.3710**$^{\triangle}$
met9	E	**0.1959**$^{\triangle}$	0.2903	0.2871

Table 1: Evaluation on the 31 TREC Terabyte topics (top10): significance tested against the baseline (set A).

judged non-relevant documents is small, but expect to obtain an increasingly good estimate using the non-relevant documents as the size of this set increases. Thus, we compare our model using explicit non-relevance information to the same model using the collection as a non-relevance model, by submitting two distinct runs: **met6**, using the set of non-relevant documents, and **met9**, using only the collection ($\delta_2 = 1$, viz. Eq. 7).

Preprocessing and Parameter settings We did not perform any preprocessing of the data besides standard stopword removal and stemming using a Porter stemmer. For our models we need to estimate four parameters: δ_1, δ_2, λ_D, and λ_Q. We have used the odd numbered topics from the TREC Terabyte track (topics 701-850) and from the TREC Million Query track (topics 1-10000) as training data. We have performed sweeps (with steps of 0.1) over possible values for these parameters and select the parameter settings with the highest resulting MAP scores. The resulting set of parameters that we have used for **met6** is given by: $\lambda_D = 0.2$, $\lambda_Q = 0.4$, $\delta_1 = 0.2$, and $\delta_2 = 0.6$. The settings for **met9** are: $\lambda_D = 0.2$, $\lambda_Q = 0.4$, and $\delta_1 = 0.2$.

6 Results and Discussion

The results of our 10 individual runs are listed in Table 1 and Table 2. Figures 1 and 2 further illustrate the differences on the 31 TREC Terabyte topics. We use the Wilcoxon signed-rank test to test for significant differences between runs and report on significant increases (or drops) for $p < .01$ using $^{\blacktriangle}$(and $^{\blacktriangledown}$) and for $p < .05$ using $^{\triangle}$(and $^{\triangledown}$). Note that the baseline runs (set A) are the same for both methods, since neither uses any kind of relevance information. The same holds for set B: in this set only relevant information is available and the two methods should therefore result in the same scores. Due to a small bug in the implementation, however, parameter δ_2 was not properly normalized, causing a slight difference in the retrieval results for **met6** on set B.

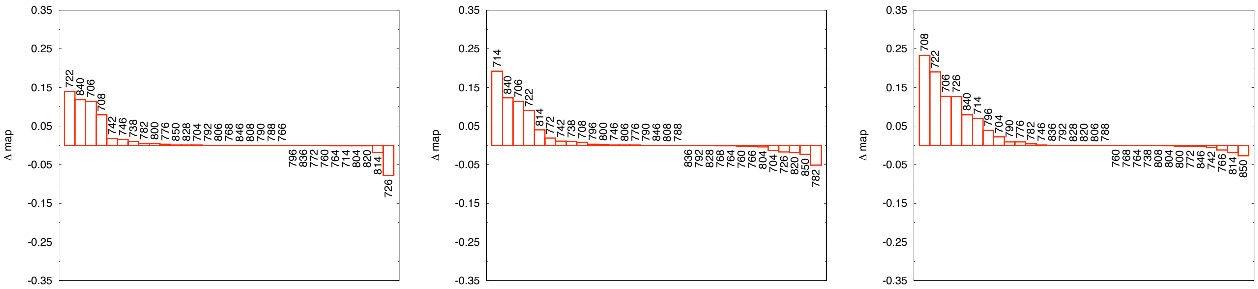
	setA	setB	setC	setD	setE
met6	0.2289	0.2595$^{\blacktriangle}$	0.2750$^{\blacktriangle}$	0.2758$^{\blacktriangle}$	**0.2822**$^{\blacktriangle}$
met9	0.2289	0.2608$^{\blacktriangle}$	0.2787$^{\blacktriangle}$	0.2777$^{\blacktriangle}$	**0.2810**$^{\blacktriangle}$

Table 2: Evaluation with statMAP: significance tested against baseline (set A).

As stated earlier, we submitted our runs to explore three main research questions:

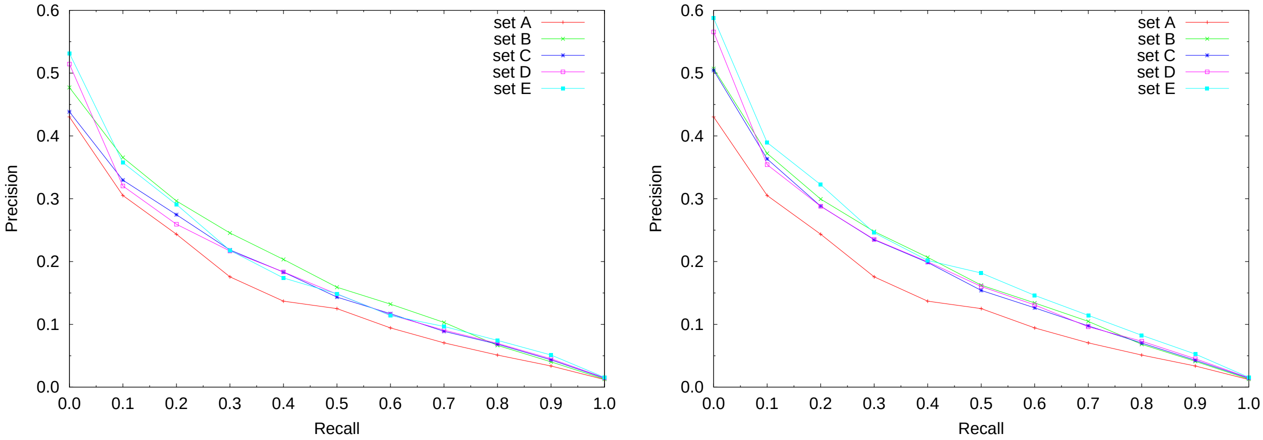
- Can non-relevance information be effectively modeled to improve the estimation of a query model?
- What is the effect of the relative size of the set of non-relevant documents with respect to the relevant documents on retrieval effectiveness?
- What are the effects when we substitute the estimates on the non-relevant documents with more general estimates, such as from the collection.

The results reported in Table 1 and Figure 2 with respect to **met6** give an answer to the first question. In all conditions, i.e., in all three measures as well as for different relevance feedback sets, the retrieval performance improves over the baseline, which confirms that our model can effectively incorporate non-relevance information for query modeling. Given a limited amount of non-relevant documents (sets C and D), our model especially improves early precision, although not significantly. A larger amount of non-relevant documents (set E) decreases overall retrieval effectiveness. From Figure 2a we observe that set E only outperforms the other sets at the very ends of the graph. Figure 1 shows a per-topic breakdown of the difference in MAP between the two submitted runs. We observe that most topics are helped more using the collection-based estimates. We have to conclude that, for the TREC Terabyte topics, the estimation on the collection yields the highest retrieval performance and is thus a better estimate of non-relevance than the judged non-relevant documents.
When we zoom out and look at the full range of available topics (Table 2), we observe that both models improve statMAP over the baseline (set A) for the full set of topics. When the feedback set is small, **met9** improves statMAP more effectively than **met6**, i.e., the background model is performing better than the non-relevant documents. On the largest set of feedback documents (set E) **met6** obtains the highest statMAP score (although the difference with **met9** is not significantly different for this set, tested using a Wilcoxon sign rank test). The difference does seem to suggest that the amount of non-relevance information needs to reach a certain size to outperform the estimation on the collection. Since we select the terms that are most likely to be sampled from the distribution of the relevant documents rather than non-relevant documents, it is crucial that the underlying relevant and non-relevant distributions can be accu-



(a) Set C (b) Set D (c) Set E

Figure 1: Per topic difference in MAP between **met6** and **met9** on the 31 TREC Terabyte topics and the various sets of relevance feedback information (a positive value indicates that **met9** outperforms **met6** and vice versa). The labels indicate the respective topic identifiers.



(a) **met6** (b) **met9**

Figure 2: Precision-recall plots of **met6** (a) and **met9** (b) on the various feedback sets and the 31 TREC Terabyte topics (top10).

rately estimated. While the relevant documents are topically concentrated, i.e., they are all related to a given query, the non-relevant documents can be topically diverse and therefore more difficult to be estimated when the number of examples is limited. The background information is generally a good approximation of the distribution of non-relevant documents, given that most of the documents in the collections are not relevant. On the other hand, as the size of the set of non-relevant examples increases, especially the query-specific top-ranked non-relevant documents, we can more accurately estimate the true distribution of the non-relevant information, which enables our model to have more discriminative power. Where this cut-off point lies remains a topic for future work.

7 Conclusion

The results presented here provide us with mixed evidence regarding the hypothesis we stipulated in Section 5. Some of the presented results (statMAP and Figure 2a) confirm the premise that, using **met6**, a larger number of judged non-relevant documents improve retrieval effectiveness most. On the other hand, the overall results obtained on the 31 TREC Terabyte topics suggest that the collection is a viable and sufficient alternative.

We would like to further explore the problem in two directions. First, we intend to investigate the impact of the available judged (non-)relevant documents and their properties with respect to the estimates on the collection. Second, given the relevance assessments, we will try to find better ways of estimating the true distribution of the (non-)relevant information within our framework. We believe that, instead of using maximum likelihood estimates, more sophisticated estimation methods may be explored and applied.

8 Acknowledgments

This research was supported by the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments (`http://www.stevin-tst.org`) under project number STE-09-12, and by the Netherlands Organisation for Scientific Research (NWO) under project numbers 017.001.190, 640.001.501, 640.002.-501, 612.066.512, 612.061.814, 612.061.815, 640.004.802, and by the Virtual Laboratory for e-Science project (`http://www.vl-e.nl`), which is supported by a BSIK grant from the Dutch Ministry of Education, Culture and Science and is part of the ICT innovation program of the Ministry of Economic Affairs.

9 References

Balog, K., Weerkamp, W., and de Rijke, M. (2008). A few examples go a long way: constructing query models from elaborate query formulations. In *SIGIR '08*.

Dunlop, M. D. (1997). The effect of accessing nonmatching documents on relevance feedback. *ACM Trans. Inf. Syst.*, 15(2):137–153.

Ide, E. (1971). New experiments in relevance feedback. In Salton, G., editor, *The SMART Retrieval System – Experiments in Automatic Document Processing*, pages 337–354. Prentice-Hall.

Kraaij, W. (2004). *Variations on Language Modeling for Information Retrieval*. PhD thesis, University of Twente.

Kurland, O., Lee, L., and Domshlak, C. (2005). Better than the real thing?: iterative pseudo-query processing using cluster-based language models. In *SIGIR '05*.

Lafferty, J. and Zhai, C. (2001). Document language models, query models, and risk minimization for information retrieval. In *SIGIR '01*.

Lavrenko, V. and Croft, B. W. (2001). Relevance based language models. In *SIGIR '01*.

Rocchio, J. (1971). Relevance feedback in information retrieval. In Salton, G., editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. Prentice Hall.

Wang, X., Fang, H., and Zhai, C. (2008). A study of methods for negative relevance feedback. In *SIGIR '08*.

Wong, W. S., Luk, R. W. P., Leong, H. V., Ho, K. S., and Lee, D. L. (2008). Re-examining the effects of adding relevance information in a relevance feedback environment. *Information Processing & Management*, In Press, Corrected Proof.

Zhai, C. and Lafferty, J. (2001). Model-based feedback in the language modeling approach to information retrieval. In *CIKM '01*.